Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 18 January 2013 (MN LATEX style file v2.2)

Maximum Likelihood Random Galaxy Catalogues and Luminosity Function Estimation

Shaun Cole[1]*
[1] *Institue for Computational Cosmology, Department of Physics, University of Durham, Science Laboratories, South Road, Durham DH1 3LE*

18 January 2013

ABSTRACT

We present a new algorithm to generate a random (unclustered) version of an magnitude limited observational galaxy redshift catalogue. It takes into account both galaxy evolution and the perturbing effects of large scale structure. The key to the algorithm is a maximum likelihood (ML) method for jointly estimating both the luminosity function (LF) and the overdensity as a function of redshift. The random catalogue algorithm then works by cloning each galaxy in the original catalogue, with the number of clones determined by the ML solution. Each of these cloned galaxies is then assigned a random redshift uniformly distributed over the accessible survey volume, taking account of the survey magnitude limit(s) and, optionally, both luminosity and number density evolution. The resulting random catalogues, which can be employed in traditional estimates of galaxy clustering, make fuller use of the information available in the original catalogue and hence are superior to simply fitting a functional form to the observed redshift distribution. They are particularly well suited to studies of the dependence of galaxy clustering on galaxy properties as each galaxy in the random catalogue has the same list of attributes as measured for the galaxies in the genuine catalogue. The derivation of the joint overdensity and LF estimator reveals the limit in which the ML estimate reduces to the standard $1/V^{\rm max}$ LF estimate, namely when one makes the prior assumption that the are no fluctuations in the radial overdensity. The new ML estimator can be viewed as a generalization of the $1/V^{\rm max}$ estimate in which $V^{\rm max}$ is replaced by a density corrected $V^{\rm dc,max}$.

Key words: galaxies: luminosity function, large-scale structure of Universe

1 INTRODUCTION

Studies of galaxy clustering as a function of the galaxy properties are placing increasingly powerful constraints on models of galaxy formation. For instance, the quantification of the dependence of the strength of galaxy clustering on luminosity and colour (Norberg et al. 2002; Zehavi et al. 2005) constrains how the distribution in mass of the dark matter halos that host the galaxies depends on luminosity and colour. This information, in turn, places very useful constraints on models of galaxy formation (e.g. Kim et al. 2009). Such techniques are being extended to new wavelengths (e.g. Guo et al. 2011) and higher redshifts (e.g. Coil et al. 2008).

Measuring the galaxy correlation function usually involves counting galaxy pairs and comparing to the expectation for an unclustered or random catalogue (Hamilton 1993; Landy & Szalay 1993). If one has a very large galaxy redshift survey then the redshift used for the random catalogue can be determined fairly accurately by fitting some assumed functional form to the observed distribution. However, this is not ideal if the survey is not large or one wants to subdivide it into smaller samples in bins of luminosity or colour. In such cases one can artificially suppress the measured clustering by over fitting random fluctuations in the redshift distribution. An alternative method is to predict the galaxy redshift distribution from an estimate of the galaxy luminosity function (LF) and the flux and other selection limits of the survey (e.g. Cole et al. 2005). The redshift distribution derived by this technique is less susceptible to distortions from density fluctuations as one can use estimators of the galaxy LF that are independent of the galaxy density (Sandage, Tammann & Yahil 1979; Efstathiou, Ellis & Peterson 1988). Also, one predicts not only the redshift, but also the luminosity of each galaxy in the random catalogue and so a single random catalogue can be used to estimate galaxy clustering as a function of luminosity. However if one wants to extend this technique so that one can measure galaxy clustering as a function of other properties, e.g. colour and surface brightness, one has the more complicated task of first estimating a multi-variate luminosity-colour-surface brightness distribution function.

We develop a new algorithm for generating a random galaxy catalogue that corresponds to a given observed catalogue defined by a simple flux limit. This is a maximum likelihood estimator for

* e-mail: shaun.cole@durham.ac.uk

the LF, $\Phi(L)$, in which, like the standard $1/V^{\max}$ (Schmidt 1968; Felten 1976) estimator, $\Phi(L)$ reduces to a weighted sum over the galaxies with luminosity L, but unlike $1/V^{\max}$ explicitly accounts for fluctuations in the galaxy density with redshift. As each observed galaxy contributes linearly to this estimated LF, this means that a random catalogue with a consistent LF can be generated by simply cloning galaxies from the observed catalogue, with a rate which we derive from a maximum likelihood analysis, and redistributing them uniformly over the volume in which they would satisfy the survey selection criteria. As each galaxy in the random catalogue is a clone of an observed galaxy it carries with it all the measured properties of that galaxy. Hence, provided they can be modified for the change in redshift (e.g. k-correcting luminosities), the resulting random catalogue has all the properties of the original and can be used to study clustering as a function of any of those properties. This technique should be particularly applicable to multi-wavelength surveys such as GAMA (Driver et al. 2011) and its overlap with H-ATLAS (Eales et al. 2010), 6dF (Jones et al. 2009), zCOSMOS (Lilly et al. 2007) and future redshift surveys designed to probe galaxy evolution.

In Section 2 we develop a joint maximum likelihood estimator for an assumed non-evolving LF and the run of overdensity as a function of redshift. We, also, show how the LF estimator relates to the standard $1/V^{\max}$ estimator. Section 3 extends this estimator to include galaxy evolution. In Section 4 we show how the estimator can be extended to provide a simple algorithm for generating a random galaxy catalogue. The method is tested and illustrated with mock data in Section 5 and we conclude in Section 6.

2 LUMINOSITY FUNCTION ESTIMATION

The commonly used STY (Sandage, Tammann & Yahil 1979) and EEP (Efstathiou, Ellis & Peterson 1988) maximum likelihood estimators of the galaxy luminosity function (LF) assume the probability of a galaxy having luminosity in the interval $L - dL/2$ to $L + dL/2$ in a volume element $d^3\mathbf{x}$ centred at position $\mathbf{x}$ can be factorized as

$$P(L, \mathbf{x})\, dL\, d^3\mathbf{x} = \phi(L)\rho(\mathbf{x})\, dL\, d^3\mathbf{x}. \tag{1}$$

They then construct estimators that are independent of the density, $\rho(\mathbf{x})$, by factoring out its dependence.

Thus they start with the following *conditional* probability

$$p_\alpha = \frac{\phi(L_\alpha)}{\int_{L^{\min}(z_\alpha)}^{\infty} \phi(L)\, dL} \tag{2}$$

that in an apparent magnitude limited catalogue a galaxy α at redshift z_α will have luminosity L_α

The STY and EEP methods differ in that STY assume a parametric (Schechter function) form for the LF, while EEP simply adopt a stepwise (binned) description of the LF. In both cases the derivation of the LF estimator follows by forming the likelihood, which is the total probability for the whole galaxy sample given the model parameters,

$$\mathcal{L} = \Pi_\alpha p_\alpha, \tag{3}$$

and maximising this likelihood (or its logarithm) over the model parameters (bin values in the case of EEP).

If we are interested in estimating both the LF and the spherically averaged density field we can instead start with the *joint* probability

$$p_\alpha = \frac{\Delta(z_\alpha)\frac{dV(z_\alpha)}{dz}\phi(L_\alpha)}{\int \Delta(z)\frac{dV}{dz}\int_{L^{\min}(z)}^{\infty} \phi(L)\, dL\, dz} \tag{4}$$

of finding a galaxy at redshift z_α with luminosity L_α in an apparent magnitude limited sample. Here dV/dz is the differential of the survey volume with redshift and $\Delta(z)$ is the galaxy overdensity (averaged over a radial bin) at redshift z. Here we are assuming that there is no redshift evolution of the luminosity function and hence $\rho(\mathbf{x})$ varies only due to density fluctuations. Adopting binned estimates of both the luminosity function ϕ_i and overdensity Δ_p we can write this probability as

$$p_\alpha = \frac{\sum_p V_p \Delta_p\, D(z_\alpha|z_p) \sum_i \phi_i\, D(L_\alpha|L_i)}{\sum_p V_p\, \Delta_p \sum_i \phi_i\, S(L^{\min}{}_p|L_i)}. \tag{5}$$

Here the sum over p (later also q) runs over redshift bins with V_p being the volume and Δ_p the galaxy overdensity of the bin. The sum over i (later also j) runs over the bins in the luminosity function with ϕ_i being equal to $\phi(L)\, dL$ for that bin. The functions $D(z_\alpha|z_p)$ and $D(L_\alpha|L_i)$ represent simple binning functions which are unity if galaxy α falls in the corresponding redshift and luminosity bin and zero otherwise. Similarly $S(L_p^{\min}|L_i)$ is a step-function which is unity if the minimum luminosity $L_p^{\min}$ required for a galaxy to make it into the magnitude limited sample at the redshift of bin p is fainter than the luminosity L_i of that bin. Using this notation we can write

$$\begin{aligned} \ln \mathcal{L} \quad = \quad & \sum_\alpha \Big(\ln \sum_p V_p \Delta_p\, D(z_\alpha|z_p) + \ln \sum_i \phi_i\, D(L_\alpha|L_i) \\ & - \ln \sum_p V_p \Delta_p \sum_i \phi_i\, S(L_p^{\min}|L_i) \Big). \end{aligned} \tag{6}$$

For the maximum likelihood solution, the derivatives of $\ln \mathcal{L}$ with respect to bin values Δ_q and ϕ_j will be zero. Hence we have

$$\begin{aligned} \frac{d \ln \mathcal{L}}{d\Delta_q} = \quad 0 \quad & = \sum_\alpha \frac{V_q\, D(z_\alpha|z_q)}{\sum_p V_p \Delta_p\, D(z_\alpha|z_p)} \\ & - \quad \sum_\alpha \frac{V_q \sum_i \phi_i\, S(L_q^{\min}|L_i)}{\sum_p V_p \Delta_p \sum_i \phi_i\, S(L_p^{\min}|L_i)} \end{aligned} \tag{7}$$

and

$$\begin{aligned} \frac{d \ln \mathcal{L}}{d\phi_j} = \quad 0 \quad & = \sum_\alpha \frac{D(L_\alpha|L_j)}{\sum_i \phi_i\, D(L_\alpha|L_i)} \\ & - \quad \sum_\alpha \frac{\sum_p V_p \Delta_p\, S(L_p^{\min}|L_j)}{\sum_p V_p \Delta_p \sum_i \phi_i\, S(L_p^{\min}|L_i)}. \end{aligned} \tag{8}$$

The meaning of the various terms in these equations can be made more explicit by adopting the following notation. Let the estimate of the number of galaxies in the survey based on the values of ϕ_i and Δ_p be

$$\hat{N}_{\rm tot} = \sum_p V_p \Delta_p \sum_i \phi_i\, S(L_p^{\min}|L_i). \tag{9}$$

Let the number of galaxies falling in each luminosity and redshift bin be N_i and N_p respectively and let

$$\hat{\bar{n}}_q = \sum_i \phi_i\, S(L_q^{\min}|L_i) \tag{10}$$

be the predicted mean galaxy number density in redshift bin q based on the estimated LF and assuming the mean density , i.e. $\Delta_q = 1$. Finally let

$$V_j^{\rm dc,max} = \sum_p \Delta_p V_p\, S(L_p^{\rm min}|L_j), \tag{11}$$

which is a *density corrected* version of the normal $V^{\rm max}$ in which the volume elements, V_p, are weighted by the estimated overdensities, Δ_p.

Using this notation we can rewrite the two constraint equations as

$$0 = \frac{N_q V_q}{V_q \Delta_q} - \frac{N_{\rm tot} V_q \hat{\bar{n}}_q}{\hat{N}_{\rm tot}} \quad \text{and} \quad 0 = \frac{N_j}{\phi_j} - \frac{N_{\rm tot} V_j^{\rm dc,max}}{\hat{N}_{\rm tot}}, \tag{12}$$

which rearrange to give the coupled equations

$$\Delta_q = \frac{N_q}{V_q \hat{\bar{n}}_q} \frac{\hat{N}_{\rm tot}}{N_{\rm tot}} \qquad \text{and} \qquad \phi_j = \frac{N_j}{V_j^{\rm dc,max}} \frac{\hat{N}_{\rm tot}}{N_{\rm tot}}. \tag{13}$$

To the extent to which the maximum likelihood model is a good description of the data $\hat{N}_{\rm tot} = N_{\rm tot}$ and so these equations simplify to quite intuitive estimators

$$\Delta_q = \frac{N_q}{V_q \hat{\bar{n}}_q} \qquad \text{and} \qquad \phi_j = \frac{N_j}{V_j^{\rm dc,max}}. \tag{14}$$

The first of these equations simply says that the estimate of the overdensity is the measured density divided by that predicted by the LF, while the second equation is equivalent to

$$\phi(L) = \sum_\alpha \frac{1}{V^{\rm dc,max}(L_\alpha)} \tag{15}$$

with the sum being over galaxies within that luminosity bin, i.e. the normal $1/V^{\rm max}$ estimator, but with $V^{\rm max}$ replaced by $V^{\rm dc,max}$.

We note that this maximum likelihood estimate of the LF is equivalent to the standard $1/V^{\rm max}$ estimator if one makes the prior assumption that $\Delta_q \equiv 1$, i.e. that there are no fluctuations in the radial galaxy density.

Choloniewski (1986) derived the same estimator of the LF using a different approach in which it was assumed that the number of galaxies in a given luminosity and redshift bin were drawn from a Poisson distribution. Our derivation shows that the estimator does not depend on the details of the assumed statistical distribution. The same density estimator was derived by maximum likelihood in section 8 of Saunders et al. (1990). They also stated that an improved estimate of the LF could be made by making the same *density correction* to $V^{\rm max}$, though they did not derive this result via maximum likelihood. Another related analysis is that of Heyl et al. (1997). They followed similar steps but choose not to make the separability assumption of equation (1) so as to be able to directly probe evolution of the shape of the LF using wide redshift bins.

Before detailing our simple algorithm for generating a random catalogue that is consistent with the LF given by equation (15), we will generalize this result to take account of redshift evolution. The resulting algorithm, described in Section 4, can then be applied to surveys that span a wide range of redshifts.

3 ALLOWING FOR REDSHIFT EVOLUTION

First let us consider the case where one has external knowledge of the evolution of the galaxy population. For instance, one might have evolutionary corrections for each galaxy or an average for the population based on fitting stellar population synthesis models (e.g. Bruzual & Charlot 2003; Blanton & Roweis 2007) to the observed galaxy colours. One could also have a pre-imposed model for density evolution, e.g. that the amplitude of the galaxy luminosity function, Φ^*, varies with redshift as $\Phi^*(z) = P(z)\Phi^*(0)$. In this case the only changes that are needed to the above estimators are:

(i) when computing the redshift range over which a given galaxy satisfies the catalogue selection criteria include the e-correction along with the k-correction and

(ii) include the factor $P(z)$, by which Φ^* evolves, in the definition of $V_\alpha^{\rm dc,max}$.

Thus, we redefine $V^{\rm dc,max}$ for galaxy α used in equation (15) to be

$$V_\alpha^{\rm dc,max} = \sum_p \Delta_p P_p V_p\, S(L_p^{\rm min}|L_\alpha), \tag{16}$$

which simply represents and integral over the survey volume weighted by the combined factor $\Delta(z)P(z)$ with limits set by the redshift range over which galaxy α would satisfy the survey selection criteria.

If one does not have foreknowledge of the evolution one can instead parameterise the evolution and use the survey data to constrain its parameters by an extension of the maximum likelihood technique. For instance for the $P(z)$ model of Φ^* evolution introduced above, equation (6) becomes

$$\begin{aligned}\ln \mathcal{L} \;=\; \sum_\alpha \Big(& \ln \sum_p V_p P_p \Delta_p\, D(z_\alpha|z_p) + \ln \sum_i \phi_i\, D(L_\alpha|L_i) \\ & - \ln \sum_p V_p P_p \Delta_p \sum_i \phi_i\, S(L_p^{\rm min}|L_i) \Big).\end{aligned} \tag{17}$$

Here the parametric form of $P(z)$ might simply be $P(z) = 1 + az$ with a being the evolution parameter we wish to determine. The method is easily generalized to more parameters. As P_p and Δ_p always appear as a pair in this likelihood function they are degenerate, i.e. we are unable distinguish evolution in the number density of galaxies from a redshift dependent change in the overdensity. If, however, we are able to specify the expected amplitude of the density fluctuations then this will enable the likelihood analysis to distinguish fluctuations from smooth evolution[1]. If the redshift bins are sufficiently large in volume we can make a simple estimate of the expected fluctuations in the galaxy overdensity using the integral $J_3 = \int \xi(r) r^2 dr$ (assumed to be a constant when integrated to scales $\gtrsim 10h^{-1}$ Mpc) of the galaxy correlation function, $\xi(r)$ (Peebles 1980). The resulting expected variance in Δ_p is

$$\sigma_p^2 = \frac{1 + 4\pi \hat{\bar{n}}_p J_3}{\hat{\bar{n}}_p V_p}, \tag{18}$$

with the second term enhancing the variance above the Poisson value because galaxy positions are correlated and tend to come in clumps of $4\pi\hat{\bar{n}}J_3$ galaxies at a time. Assuming the density fluctuations are Gaussian distributed with this variance and including this as a prior probability which multiplies our likelihood function, $\mathcal{P} = \mathcal{L} \times \mathcal{P}_{\rm prior}$, we can replace equation (17) with the following equation for the logarithm of the posterior probability (to within an unimportant additive constant)

$$\ln \mathcal{P} = \sum_\alpha \Big(\ln \sum_p V_p P_p \Delta_p\, D(z_\alpha|z_p) + \ln \sum_i \phi_i\, D(L_\alpha|L_i)$$

[1] If the estimate of the variance of the density fluctuations is inaccurate or the function $P(z)$ is given too much freedom then this may lead to bias in the recovered evolution parameters, but for the smooth evolution models considered here we find no evidence of bias.

$$- \ln \sum_p V_p P_p \Delta_p \sum_i \phi_i \, S(L_p^{\min}|L_i) \Big) - \sum_p \frac{(\Delta_p - 1)^2}{2\sigma_p^2}. \quad (19)$$

The final term breaks the degeneracy between P_p and Δ_p and so allows us to solve for the evolution parameter. In some instances, e.g. for a small survey in which density evolution is inevitably poorly constrained, it may be beneficial to place a Gaussian prior

$$\mathcal{P}_{\rm prior}(a) = \frac{1}{\sqrt{2\pi}\sigma_a} \exp(-a^2/2\sigma_a^2) \quad (20)$$

on the density evolution parameter.

The final modification is to use a Lagrange multiplier, μ, to impose the constraint that, in the absence of density fluctuations, the predicted number of galaxies, $\sum_q \hat{\hat{n}}_q V_q$, equals the number in the genuine catalogue, $N_{\rm tot}$. In the simple case presented in Section 2 this is not necessary as the likelihood expression of equation (6) is invariant under the transformation $\phi_i \rightarrow \theta\phi_i$ and $\Delta_p \rightarrow \Delta_p/\theta$. Thus, in that case one can simply impose this normalization constraint after having found the ML solution. However, the introduction of last term in equation (19) has broken this symmetry and so to maximise equation (19) subject to this constraint we need instead to maximise

$$\ln \Lambda = \ln \mathcal{P} - \mu \sum_q (\hat{\hat{n}}_q V_q - N_{\rm tot}). \quad (21)$$

Following the same steps that led from equation (6) to (12), but now also setting the derivatives

$$\frac{d \ln \Lambda}{da} = 0 \qquad \text{and} \qquad \frac{d \ln \Lambda}{d\mu} = 0, \quad (22)$$

where μ is the Lagrange multiplier and a is the parameter of the evolution model $P(z)$, leads to the following ML solution,

$$0 = \frac{N_q}{\Delta_q} - V_q \hat{\hat{n}}_q - \frac{\Delta_q - 1}{\sigma_q^2} \quad (23)$$

$$0 = \frac{N_j}{\Phi_j} - \left(V_j^{\rm dc,max} + \mu V_j^{\rm max}\right) \quad (24)$$

$$0 = \sum_q \left(N_q - \hat{\hat{n}}_q V_q (\Delta_q + \mu)\right) \frac{d \ln P_q}{da} - \frac{a}{\sigma_a^2} \quad (25)$$

$$0 = \sum_q \hat{\hat{n}}_q V_q - N_{\rm tot}. \quad (26)$$

Here we have generalized the earlier notation to include the $P(z)$ model so that

$$\hat{\hat{n}}_q = P_q \sum_i \phi_i \, S(L_q^{\min}|L_i), \quad (27)$$

$$V_j^{\rm dc,max} = \sum_p \Delta_p P_p V_p \, S(L_p^{\min}|L_j), \quad (28)$$

and made use of the result that if the model accurately describes the data then

$$\hat{N}_{\rm tot} = \sum_p P_p V_p \Delta_p \sum_i \phi_i \, S(L_p^{\min}|L_i) = N_{\rm tot}. \quad (29)$$

These equations can be solved efficiently by an iterative method. Starting with $\Delta_q \equiv 1$ and $P_q \equiv 1$ (or a prior guess for the evolution parameter a).

(i) Evaluate $V^{\rm dc,max}$ and $V^{\rm max}$ for each galaxy using the current values of Δ_q and P_q.

(ii) Find the value of μ such that $\left\langle \frac{V_\alpha^{\rm max}}{V_\alpha^{\rm dc,max} + \mu V_\alpha^{\rm max}} \right\rangle = 1$, which is achieved easily using the Newton-Raphson method.

(iii) Evaluate $\hat{\hat{n}}_q$ using

$$\hat{\hat{n}}_q V_q = \sum_\alpha \frac{P_q V_q S(L_q^{\min}, L_\alpha)}{V_\alpha^{\rm max}} \left(\frac{V_\alpha^{\rm max}}{V_\alpha^{\rm dc,max} + \mu V_\alpha^{\rm max}} \right), \quad (30)$$

which follows from evaluating equation (27) using the estimate of ϕ_j given by equation (24).[2]

(iv) Substitute this estimate of $\hat{\hat{n}}_q$ into equation (23) to solve for the Δ_q.

(v) Solve for the number density evolution parameter, a, by finding the root of equation (25).[3]

(vi) Now repeat this process from step (i) until the Δ_q and the P_q converge.

In the iterative process described above we never explicitly evaluate the luminosity function, $\Phi(L)$, though one could do this at any stage by simply evaluating

$$\phi(L) = \sum_\alpha \frac{1}{V^{\rm dc,max}(L_\alpha) + \mu V^{\rm max}(L_\alpha)}, \quad (31)$$

which follows from equation (24). Hence although we derived the method by considering a binned estimate of the luminosity function this binning does not enter in any way in determining the parameters Δ_q and a or into the predicted redshift distribution, $\hat{\hat{n}}_q V_q$, they imply.

One could deal with luminosity evolution in an analogous way. First define the e-correction term in the standard way so that absolute, M, and apparent, m, magnitudes are related by

$$M = m - 5 \log_{10} d_{\rm lum}(z) - k(z) - e(z), \quad (32)$$

where $d_{\rm lum}$ is the luminosity distance and $k(z)$ the k-correction (see e.g. Hogg et al. 2002). Then parameterize the e-correction (or its deviation from a default individual e-correction for each galaxy) as e.g. $e(z) = uz$ and maximize the posterior probability with respect to the parameter u. This yields the constraint equation

$$\frac{d \ln \mathcal{P}}{du} = 0 = \sum_j \frac{dN_j}{du} \ln \Phi(L_j) - \sum_p V_p P_p (\Delta_p + \mu) \phi(L_p^{\min}) \frac{dL_p^{\min}}{du} - \frac{u}{\sigma_u^2}, \quad (33)$$

where the last term comes from assuming a Gaussian prior on the evolution parameter. The other terms depend on u through the implicit dependence of the luminosities L_α and $L_p^{\min}$ on the e-correction via the relationship between the inferred absolute magnitude, the observed apparent magnitude, m_α and redshift z_α,

$$M_\alpha = m_\alpha - 5 \log_{10} d_{\rm lum}(z_\alpha) - k(z_\alpha) - e(z_\alpha), \quad (34)$$

[2] We have written the equation in this form as if we then sum over the redshift bins, q, it is straightforward to see that the choice of μ from step (ii) ensures that equation (26) is satisfied. In practice, we find $|\mu| \ll 1$ and that setting $\mu = 0$ makes very little difference to the resulting LF and redshift distribution.

[3] Here we assume that as $\hat{\hat{n}}_q \propto P_q / \sum_q \hat{\hat{n}}_q V_q$, which is appropriate if ϕ_j and Δ_q are being held fixed and the normalization constraint, equation (26), is being maintained. The approximate scaling of $\hat{\hat{n}}_q$ used in step (v) does not have to be exact. We use it as a fast way of estimating $\hat{\hat{n}}_q$ at any value of the evolution parameter a from the existing estimate we have at $a = a'$ from step (iii). Once we have iterated these equations to the point they converge then $a \approx a'$ and so these scaling factors all tend to unity. The approximation used in this scaling only effects the speed of convergence.

and through the dependence of the limiting absolute magnitude at redshift z_p on the apparent magnitude limit of the survey, $m_{\rm faint}$,

$$M_{\rm faint} = m_{\rm faint} - 5\log_{10} d_{\rm lum}(z_p) - k(z_p) - e(z_p). \tag{35}$$

Hence, u can be found in an iterative way, updating u by finding the root of equation (33) in the same way as we update a by finding the root of equation (25). Implementing this modified alogrithm requires a smooth luminosity binning scheme, as in Efstathiou et al. (1988), so that the derivative dN_j/du is well defined. Although we have successfully implemented such a scheme we prefer to present results in which we use the simpler iterative algorithm detailed above. This is sufficiently fast that we can repeat it for different fixed values of the e-correction (u), iterating to the final solution for each value of u, and then search over the values of u to find the value which maximises the logarithm of the posterior probability

$$\begin{aligned}\ln\mathcal{P} &= \sum_\alpha \Big(\ln\sum_p V_p\Delta_p\, D(z_\alpha|z_p) + \ln\sum_i \phi_i\, D(L_\alpha|L_i) \\ &\quad - \ln\sum_p V_p\Delta_p \sum_i \phi_i\, S(L_p^{\rm min}|L_i) \Big) \\ &\quad - \sum_p \frac{(\Delta_p-1)^2}{2\sigma_p^2} - \frac{a^2}{2\sigma_a^2} - \frac{u^2}{2\sigma_u^2}. \end{aligned} \tag{36}$$

The initial terms come from equation (19) and the terms on the final line of this equation come from assumed Gaussian priors on the evolution parameters a and u. The term on the second line is effectively constant as it involves only the total number of galaxies predicted by the model. Thus, to within an unimportant additive constant we can evaluate this expression as

$$\begin{aligned}\ln\mathcal{P} &= \sum_\alpha \Big(\ln\sum_p V_p\Delta_p\, D(z_\alpha|z_p) + \ln\sum_i \phi_i\, D(L_\alpha|L_i) \Big) \\ &\quad - \sum_p \frac{(\Delta_p-1)^2}{2\sigma_p^2} - \frac{a^2}{2\sigma_a^2} - \frac{u^2}{2\sigma_u^2}, \end{aligned} \tag{37}$$

or equivalently in terms of the binned quantities as

$$\begin{aligned}\ln\mathcal{P} &= \sum_p N_p \ln(V_p\Delta_p) + \sum_i N_i \ln(\phi_i) \\ &\quad - \sum_p \frac{(\Delta_p-1)^2}{2\sigma_p^2} - \frac{a^2}{2\sigma_a^2} - \frac{u^2}{2\sigma_u^2}. \end{aligned} \tag{38}$$

Thus for each trial value of the luminosity evolution parameter u one evaluates this expression using the values of ϕ_i and Δ_p that result from the iterative solution of equations (23) to (26) and then simply selects the most probable model.

4 GENERATING A RANDOM CATALOGUE

The LF estimates we have derived in Sections 2 and 3 are both simply weighted sums over the galaxies of that luminosity. This feature means they are very well suited for generating random catalogues. Rather than having to estimate the LF and then compute the number of galaxies expected at a given redshift in the random catalogue as an integral over $\Phi(L)$, one can instead carry out a weighted duplication of the galaxies in the original catalogue with each being redistributed in redshift.

The key to the algorithm is equation (30). The left hand side of this equation is the predicted number of galaxies in the redshift bin z_q of the random catalogue. The right hand side of the equation we can interpret as saying each galaxy in the original catalogue has a weight $w_\alpha = \frac{V_\alpha^{\rm max}}{V_\alpha^{\rm dc,max}+\mu V_\alpha^{\rm max}}$ and because $V_\alpha^{\rm max} \equiv \sum_q P_q V_q S(L_q^{\rm min}, L_\alpha)$ we see that the first term indicates that this weight is distributed amongst the redshift bins according to the fraction of its $V^{\rm max}$ that falls within each bin. This interpretation of equation (30) leads to a very simple Monte Carlo algorithm for generating a random catalogue, i.e. the galaxy catalogue one would expect if there were no galaxy clustering.

To generate a random catalogue with approximately $N_{\rm times}$ as many galaxies as the original we proceed as follows. Loop over the galaxies in the original catalogue and, for each one, place $N_{\rm times}w_\alpha$ duplicates[4] into the random catalogue, with the redshift of each duplicate being randomly selected within the volume $V^{\rm max}$ that is accessible to that galaxy. These weights correct for the fact that galaxies of a given luminosity may be over- or under-represented in the original catalogue as a result of density fluctuations within the volume probed by the catalogue. The definition of $V^{\rm max}$ used here should include the $P(z)$ factor, but not $\Delta(z)$, i.e.

$$V^{\rm max}(z_\alpha^{\rm max}) = \int_0^{z_\alpha^{\rm max}} \frac{dV}{dz} P(z) dz, \tag{39}$$

where $z_\alpha^{\rm max}$ is the redshift at which the galaxy α would drop outside the survey selection criteria. A fast algorithm to achieve this is to first generate a lookup table for $V^{\rm max}(z)$. Then, for the clone of each galaxy, α, one generates a uniform random variable, s, in the interval $[0, 1]$ and uses the lookup table to assign it the redshift at which $V^{\rm max}(z) = sV^{\rm max}(z_\alpha^{\rm max})$. The redshift dependent properties of the galaxy such as apparent magnitude must be adjusted using the distance modulus, k- and e-corrections to this assigned redshift. The angular position of the galaxy can be independently randomly chosen within the angular footprint of the survey. The result is a random catalogue with a smooth redshift distribution and luminosity function consistent with the maximum likelihood value given by equation (31).[5]

5 RESULTS

As a first test of our algorithm we have analysed a mock galaxy catalogue that has been constructed from the Virgo Millennium Simulation (Springel et al. 2005). The simulation was populated with galaxies using the Bower et al. (2006) version of the GALFORM semi-analytic model.[6]

In Fig. 1 we show the redshift distribution of a shallow, $r < 17.5$ and $z < 0.2$, portion of a 1000 square degree region of this mock catalogue. The redshift distribution is very structured as a result of realistic large scale structure – voids, filaments and clusters – in the three dimensional galaxy distribution (Springel et al. 2005). The smooth curves in the upper panel of Fig. 1 show the redshift distributions of our corresponding random catalogues. The dotted

[4] Although this ratio is not in general an integer one can round up or down with probabilities chosen such that the mean is the required value.

[5] A related random catalogue algorithm was explored in Cresswell (2010), but without applying the density dependent weights, w_α that are required by this maximum likelihood derivation. Cresswell (2010) used the resulting redshift distribution as an alternative to LF based prediction employed in Cresswell & Percival (2009) when quantifying scale dependent bias for red and blue galaxies in SDSS.

[6] This catalogue is a prototype of set of mock Pan-STARRS galaxy catalogues available at https://ps1-durham.dur.ac.uk/mocks.

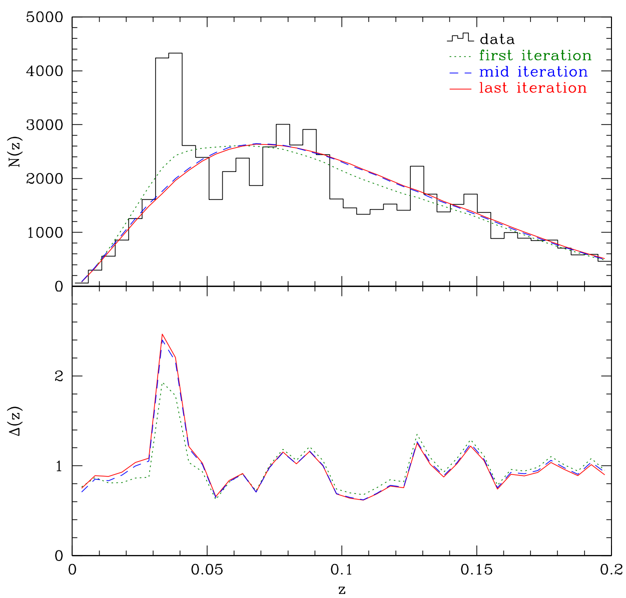

Figure 1. The upper panel compares the redshift distribution of the data from a mock catalogue with the predicted smooth redshift distributions of selected iterations of the random catalogue. The first iteration is shown by the dotted (green) curve and a subsequent and final iteration by the dashed (blue) and solid (red) curves respectively. The lower panel shows the overdensity in redshift shells, $\Delta(z)$, of the mock catalogue compared to the different iterations of the random catalogue. In both panels the dashed (blue) curves are almost coincident with the solid (red) curves.

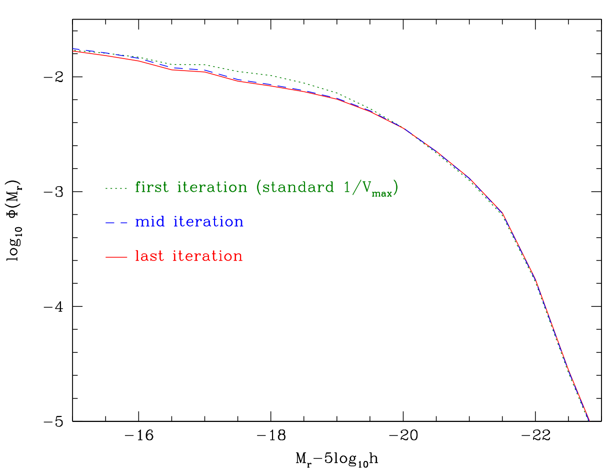

Figure 2. The r-band luminosity function of selected iterations of the random catalogue. The estimate from the first iteration, shown by the dotted (green) curve, is simply the standard $1/V^{\max}$ estimate of the luminosity function. Subsequent iterations, shown by the dashed (blue) and almost coincident solid (red) curves, rapidly converge.

(green) curve is the result of the simple algorithm in which the catalogue galaxies are just randomized within the accessible volume, $V^{\max}$, within which the galaxy could be detected and meet the selection criteria of the catalogue. In this process a simple r-band k-correction,

$$k(z) = 0.87z + 1.38z^2, \tag{40}$$

was assumed for all galaxies, this being typical of the k-correction given by Blanton & Roweis (2007) for r-band selected galaxies in the SDDS survey. The evolution, e-correction, was assumed to be negligible. Even without reference to the other models it is clear that this redshift distribution has been biased by the presence of large scale structure. For instance the overdensity at $z \approx 0.04$ results in a shoulder in the redshift distribution of the random catalogue.

The two remaining and almost identical curves in the upper panel of Fig. 1 show the redshift distributions of the random catalogues that result from taking the $V^{\max}$ based estimate as a starting point and applying the iterative procedure described in Section 4 to find the solutions to equations (14). The same k-correction and no evolution were assumed as in the $V^{\max}$ based estimate. This procedure rapidly converges to a stable random catalogue with a smooth redshift distribution which is unbiased by the large scale structure. The lower panel of Fig. 1 shows the overdensity of the mock catalogue as a function of redshift, estimated as the ratio of the redshift distribution of the mock catalogue to that of the random catalogue. It is clear that the $V^{\max}$ based estimate, like methods which simply fit the observed redshift distribution, underestimates the true amplitude of the density fluctuations and would lead to biased estimates of galaxy correlation functions and other large scale structure statistics.

The estimated luminosity functions corresponding to these different random catalogues are shown in Fig. 2. We again see excellent convergence in estimates resulting from our iterative procedure. In this case, the $1/V^{\max}$ estimate, which is our starting point, is biased high at intermediate magnitudes by the overdensity at $z \approx 0.04$.

To test the method further we set up a deeper galaxy catalogue with a known luminosity function and explicit luminosity and density evolution. To achieve this we first set up a galaxy catalogue with no spatial clustering by sampling the evolving Schechter luminosity function

$$\Phi(L) = P(z)\, \Phi_* \left(\frac{L}{L_*(z)}\right)^{-\alpha} \exp\left(-\frac{L}{L_*(z)}\right) \tag{41}$$

in a standard flat cosmology with density parameter $\Omega_{\rm m} = 0.3$ and cosmological constant $\Omega_\Lambda = 0.7$. For the parameters of this evolving Schechter function we adopted $\Phi_* = 1.49 \times 10^{-2}\, h^3\,{\rm Mpc}^{-3}$, $\alpha = 1.05$, $P(z) = \exp(0.18z)$ and $L_*(z)$ equivalent to characteristic r-band absolute magnitude $M_* = -20.37$ at $z = 0$ with an assumed e-correction term

$$e(z) = -1.62z. \tag{42}$$

The k-correction was again given by equation (40). These choices are compatible with the parameterization of the SDSS r-band luminosity estimated by Blanton et al. (2003), though they chose to work in a magnitude system referenced to $z = 0.1$. The resulting redshift distribution for an $r < 24$ magnitude limited catalogue of 5 square degrees is shown by the dashed (blue) line in the upper panel of Fig. 3, labelled "truth".

To impose density fluctuations on the smooth redshift distribution we divided the catalogue into redshift bins, with volumes V_p, and for each bin generated a random density perturbation $\delta_p > -1$ drawn from a truncated Gaussian with variance $4\pi J_3/V_p$. Here we chose $4\pi J_3 = 5000$, which is appropriate for L_* galaxies (Hawkins et al. 2003). We then generated the catalogue with the redshift distribution shown by the histogram in Fig. 3 by randomly accepting galaxies from a D times denser version of original unclustered catalogue with probability $(1+\delta_p)/D$. The Poisson fluc-

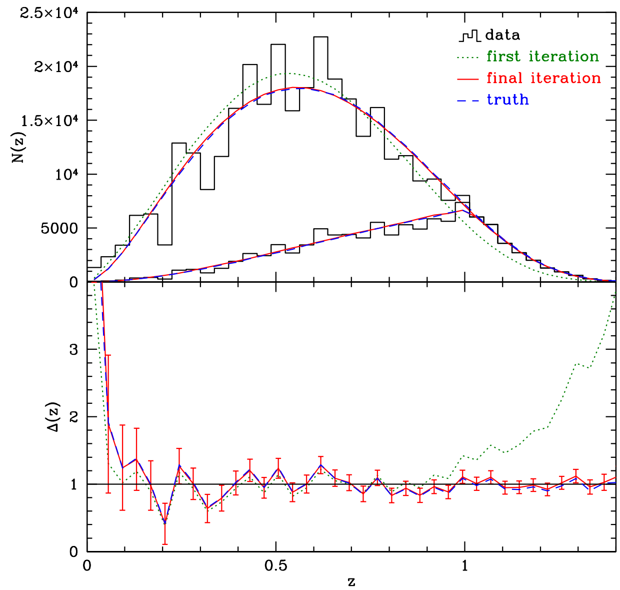

Figure 3. The upper panel shows two sets of redshift distributions. The upper distributions are for the full population of galaxies in a 5 square degree, $r < 24$ magnitude limited survey. The lower distributions are for the subset of these galaxies with absolute magnitudes $M_r < -20$. In both cases the clumpy distribution (black histograms) from the synthetic catalogue is compared with the smooth redshift distributions of two random catalogues and that of the original uniform catalogue (blue dashed curves) from which it was constructed. As described in the text the synthetic catalogue includes both luminosity and density evolution. The lower panel shows the ratio, $\Delta(z)$, of the full redshift distribution of the data to each of the random catalogues. The random catalogue shown by the dotted (green) curves, the starting point of the iterative process, is based on the $V^{\rm max}$ of each galaxy and ignores both luminosity and density evolution. For the random catalogue shown by the solid (red) curves, the iterative procedure described in Section 3 has been applied to determine the luminosity and density evolution parameters that maximise the posterior probability, equation (38). In each case the solid (red) curves are almost coincident with the (blue) dashed curves. The error bars shown in the lower panel are the expected level of fluctuations as given by equation (18).

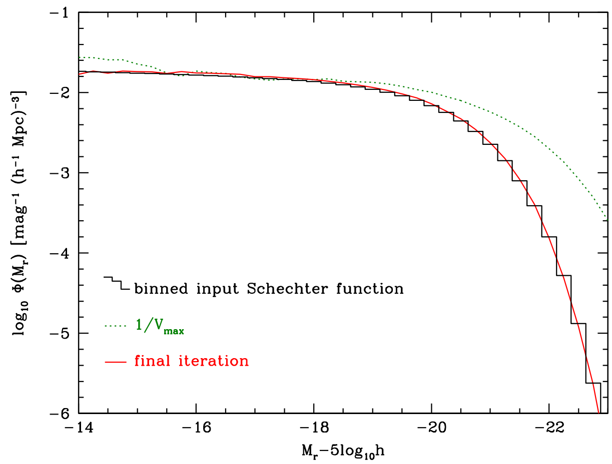

Figure 4. Comparison of the input Schechter luminosity function with those recovered by $V^{\rm max}$ and the iterative maximum likelihood method. For a fair comparison, the input Schechter function has been averaged over the 0.25 mag width bins used in the other estimates.

tuations from this sampling process combine with the imposed fluctuations, δ_p, to produce fluctuations consistent with the variance given by equation (18).

Taking this catalogue as input we generated corresponding random catalogues by applying the iterative procedure described in Section 3 to find the solutions to equations (23) to (25) and maximise the posterior probability given in equation (38). Here we assumed the density evolution to be of the form

$$P(z) = \exp((a + 0.18)z) \tag{43}$$

and the luminosity evolution of the form

$$e(z) = -0.5z + uz \tag{44}$$

with a and u being free parameters. Hence we would hope to find $a \approx 0.0$ and $u \approx -1.12$.

As the starting point of the iterative process we assumed $a = 0$ and $u = 0$ (i.e. the default density evolution, but insufficient luminosity evolution), with Gaussian priors of width $\sigma_a = 0.05$ and $\sigma_u = 1.5$. Under these assumptions the initial $V^{\rm max}$ based estimate results in a random catalogue with the redshift distribution shown by the dotted (green) curve in Fig. 3. This can be seen to be biased high at $z \lesssim 0.1$ by a local overdensity and to underpredict the number of galaxies at $z \gtrsim 0.8$ due to its lack of evolution. This is seen more clearly in the lower panel which plots the overdensity estimated as the ratio of the redshift distributions of the input catalogue to the random catalogue.

The maximum likelihood random catalogue is shown by the solid (red) curves in Fig. 3. The converged result for the evolution parameters is $a = 0.05$ and $u = -1.11$, which are close to the true values. One does not expect to recover the exact input values as the density fluctuations introduce noise into the estimates. One could determine formal errors on all the model parameters by determining the Fisher matrix from the second derivatives of the likelihood function. However, it is probably simpler, more convenient and more robust to determine the errors by repeating the whole procedure on jackknife samples of the original catalogue. For a catalogue of this particular size and depth it turns out that the density evolution parameter a is only weakly constrained and hence the prior on a is playing a role (i.e. a broader prior leads to a different a, but the resulting random catalogues are hardly distinguishable). In contrast, the luminosity evolution parameter, u, is tightly constrained and the input value is recovered quite accurately. This is true provided that sufficiently narrow magnitude bins are used for the LF. We have found that using wide bins leads to an underestimate of the degree of luminosity evolution, Broadening the underlying luminosity function by the bin width artificially boosts the bright end of the LF and so, just like luminosity evolution, it makes a tail of high redshift luminous galaxies more probable. With magnitude bins of width less than 0.5 magnitudes this effect is very small.

In Fig. 3, one can see that this procedure has produced a smooth redshift distribution that is in accurate agreement with the true underlying redshift distribution from which the synthetic catalogue was constructed. The redshift distributions that are shown in Fig. 3 for the subset of galaxies with absolute magnitudes $M_r < -20$ illustrate that the random catalogue we have produced can be

used to model the underlying smooth redshift distribution of any selected subset of the data.

We compare the input and recovered $z = 0$ luminosity functions in Fig 4. We see the initial $1/V^{\rm max}$ is shifted towards bright magnitudes due to the incorrect luminosity function and is also biased high at the faintest magnitudes due to the local $z < 0.1$ overdensity. The maximum likelihood/maximum posterior probability estimate has recovered the input luminosity function very accurately.

6 CONCLUSIONS

We have presented a maximum likelihood estimator for the galaxy luminosity function which can be viewed as an extension to the $1/V^{\rm max}$ method (Schmidt 1968), taking into account the effect of density fluctuations within the volume probed by the galaxy catalogue. The standard $V^{\rm max}$ is replaced by a *density corrected* version, $V^{\rm dc,max}$, that explicitly corrects for the over- or under-representation of galaxies of a particular luminosity in the catalogue produced by large scale structure. The utility of our luminosity function estimator is that it is a very simple and intuitive modification of the much used, but biased, $1/V^{\rm max}$ method. Similar density corrections to $1/V^{\rm max}$ have been utilised by Croton et al. (2005) and Baldry et al. (2006) to study the dependence of galaxy properties on environment and to probe the very low mass end of the stellar mass function (Baldry et al in preparation), but they used an external volume limited galaxy sample as the density defining-population rather than computing the overdensity via maximum likelihood.

We extended the maximum likelihood analysis to include arbitrary parametric models of the redshift evolution of both the characteristic luminosity and number density of the galaxy population and described a fast iterative scheme to solve the resulting equations.[7] Our analysis assumes a redshift catalogue which is complete to a single specified apparent magnitude limit. The method can be extended to include a model of magnitude dependent incompleteness by incorporating an incompleteness term into the likelihood function (e.g. see Heyl et al. 1997). To determine $V^{\rm dc,max}$ one merely needs to be able to determine over what range of redshift a given observed galaxy would continue to satisfy the survey selection criteria. Hence, in principle, it ought to possible to extend the method to surveys with colour selection. However, more work is required to see if modelling colour evolution will prove to be a barrier to getting sufficiently accurate models of such selection functions.

In both the simple and more generalized versions the estimate of the galaxy luminosity function, $\Phi(L)$, is a simple weighted sum over the galaxies of luminosity L. One consequence of this is that we have been able to specify a simple algorithm to generate unclustered, random galaxy catalogues consistent with this luminosity function by simply cloning galaxies (with a frequency determined by the weight) from the original catalogue and redistributing them uniformly throughout the survey volume in which they would be detected. At no point in this process is there any binning by luminosity and so no assumptions are required about the form or smoothness of the luminosity function. One specifies redshift bins, within which to estimate the radial overdensity, but the bin widths only very weakly affect the resulting redshift distribution of the random catalogue which is smooth and continuous. Random galaxy catalogues are widely employed when making estimates of galaxy clustering. Often used alternatives such as simple parametric fits to the observed redshift distribution are inferior as they do not use the full information available in the galaxy catalogue and are prone to either over fitting density fluctuations or failing to capture the true shape of the selection function. These shortcomings can lead to underestimating the strength of clustering on intermediate scales and overestimating the strength on the largest scales. A particular advantage of these new random catalogues is that each galaxy they contain carries with it all the measured properties that existed for the observed galaxy from which it was cloned. Hence, we expect random catalogues produced by this maximum likelihood technique to be particularly valuable for studies of how galaxy clustering depends on galaxy properties such as colour, surface brightness, morphology or spectral features.

[7] A fully documented Fortran95 subroutine that implements this algorithm and generates the related random catalogue is available at http://astro.dur.ac.uk/~cole/publications.html#software.

ACKNOWLEDGEMENTS

I would like to thank Jim Cresswell as one topic we discussed in his PhD viva was partially responsible for prompting me to revisit some unfinished work on luminosity function estimation that I had started years earlier. I thank Carlos Frenk for several useful discussions and comments on the manuscript. I also gratefully acknowledge the support of a Leverhulme Research Fellowship. This work was supported in part by an STFC rolling grant to the ICC. The calculations for this paper were performed on the ICC Cosmology Machine, which is part of the DiRAC Facility jointly funded by STFC, the Large Facilities Capital Fund of BIS, and Durham University.

REFERENCES

Baldry I. K., Balogh M. L., Bower R. G., Glazebrook K., Nichol R. C., Bamford S. P., Budavari T., 2006, MNRAS, 373, 469
Blanton M. R. et al., 2003, ApJ, 592, 819
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, MNRAS, 370, 645
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Choloniewski J., 1986, MNRAS, 223, 1
Coil A. L. et al., 2008, ApJ, 672, 153
Cole S. et al., 2005, MNRAS, 362, 505
Cresswell J. G., 2010, Portsmouth PhD. thesis
Cresswell J. G., Percival W. J., 2009, MNRAS, 392, 682
Croton D. J. et al., 2005, MNRAS, 356, 1155
Driver S. P. et al., 2011, MNRAS, 413, 971
Eales S. et al., 2010, PASP, 122, 499
Efstathiou G., Ellis R. S., Peterson B. A., 1988, MNRAS, 232, 431
Felten J. E., 1976, ApJ, 207, 700
Guo Q. et al., 2011, MNRAS, 193
Hamilton A. J. S., 1993, ApJ, 417, 19
Hawkins E. et al., 2003, MNRAS, 346, 78
Heyl J., Colless M., Ellis R. S., Broadhurst T., 1997, MNRAS, 285, 613
Hogg D. W., Baldry I. K., Blanton M. R., Eisenstein D. J., 2002, ArXiv: astro-ph/0210394
Jones D. H. et al., 2009, MNRAS, 399, 683

Kim H., Baugh C. M., Cole S., Frenk C. S., Benson A. J., 2009, MNRAS, 400, 1527
Landy S. D., Szalay A. S., 1993, ApJ, 412, 64
Lilly S. J. et al., 2007, ApJS, 172, 70
Norberg P. et al., 2002, MNRAS, 332, 827
Peebles P. J. E., 1980, The large-scale structure of the universe. Princeton University Press, 1980.
Sandage A., Tammann G. A., Yahil A., 1979, ApJ, 232, 352
Saunders W., Rowan-Robinson M., Lawrence A., Efstathiou G., Kaiser N., Ellis R. S., Frenk C. S., 1990, MNRAS, 242, 318
Schmidt M., 1968, ApJ, 151, 393
Springel V. et al., 2005, Nature, 435, 629
Zehavi I. et al., 2005, ApJ, 630, 1